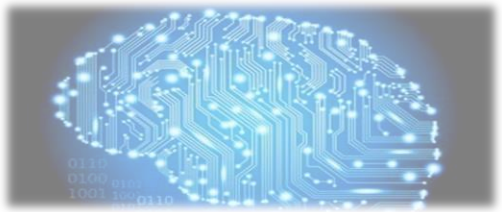
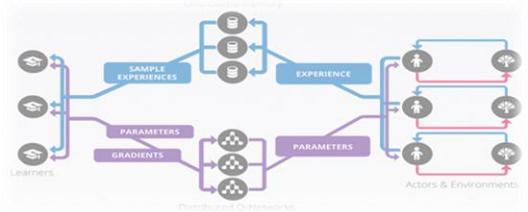



Deep Multimodal Fusion by Channel Exchanging

AutoML 연구실
석사과정 문아성

21.10.26

Outline

- 개인 연구
  - “Deep Multimodal Fusion by Channel Exchanging” 논문 리뷰
  - “Scene Text Recognition” survey

- 인공지능 대학원 지원사업
  - 산학연-3, 산학연-5 프로젝트 진행 상황

- 담낭 용종 진행 상황 보고

Traditional multimodal fusion methods

- *Aggregation-based fusion methods: To combine multimodal sub-networks into a single network*
  ex) averaging, concatenation, self-attention

$$\hat{\boldsymbol{y}}^{(i)} = f(\boldsymbol{x}^{(i)}) = h(\mathrm{Agg}(f_1(\boldsymbol{x}_1^{(i)}), \cdots, f_M(\boldsymbol{x}_M^{(i)})))$$

$$\boldsymbol{x}^{(i)} = \{\boldsymbol{x}_m^{(i)} \in \mathbb{R}^{C\times(H\times W)}\}_{m=1}^{M}$$

m: modality

h: global network

Agg: aggregation function

- *Alignment-based fusion methods: To align the embedding of all sub-networks while maintaining full propagation for each sub-network by adopting regulation loss*

$$\min_{f_{1:M}} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(\sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}_m^{(i)}), \boldsymbol{y}^{(i)}\right) + \mathrm{Alig}_{f_{1:M}}(\boldsymbol{x}^{(i)}), \quad s.t. \sum_{m=1}^{M} \alpha_m = 1$$

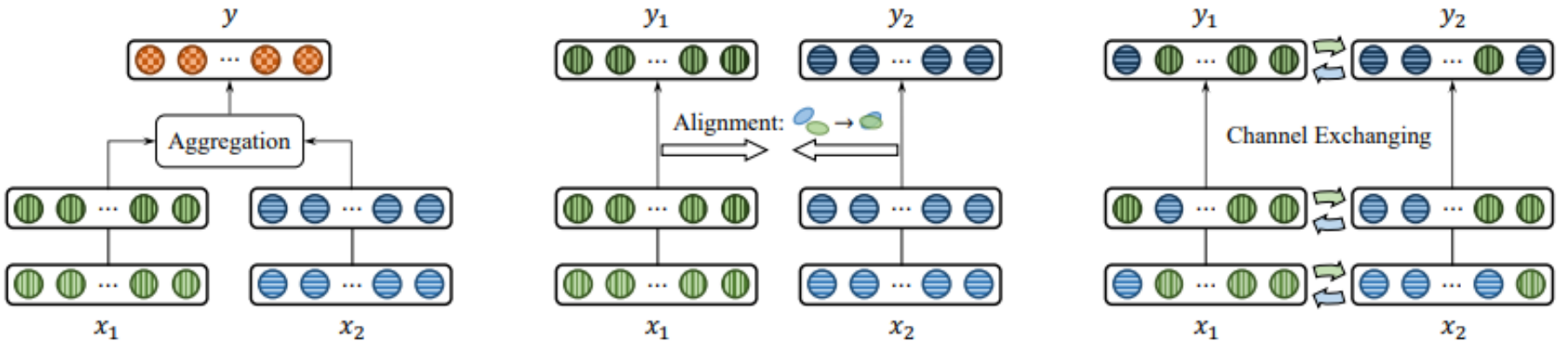
f_m: ensemble

α_m: decision score

$Alig$: loss function

Problems of traditional methods

- *Aggregation-based fusion is prone to underestimating the intra-modal propagation once the multimodal sub-networks have been aggregated.*
- *Alignment-based fusion delivers ineffective inter-modal fusion owing to the weak message exchanging by solely training the alignment loss.*



(a) Aggregation-based fusion **(b)** Alignment-based fusion **(c)** Ours

Figure 1: A sketched comparison between existing fusion methods and ours.

Channel Exchanging by Comparing BN Scaling Factor

$$\boldsymbol{x}'_{m,l,c} = \begin{cases} \gamma_{m,l,c} \frac{\boldsymbol{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma^2_{m,l,c} + \epsilon}} + \beta_{m,l,c}, & \text{if} \quad \gamma_{m,l,c} > \theta; \\ \frac{1}{M-1} \sum_{m' \neq m}^{M} \gamma_{m',l,c} \frac{\boldsymbol{x}_{m',l,c} - \mu_{m',l,c}}{\sqrt{\sigma^2_{m',l,c} + \epsilon}} + \beta_{m',l,c}, & \text{else;} \end{cases}$$

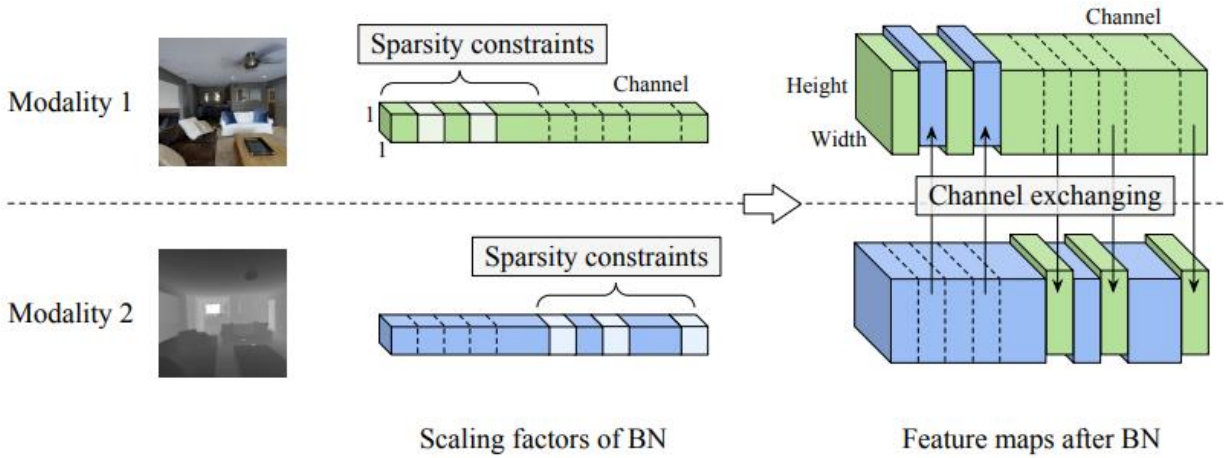


Figure 2: An illustration of our multimodal fusion strategy. The sparsity constraints on scaling factors are applied to disjoint regions of different modalities. A feature map will be replaced by that of other modalities at the same position, if its scaling factor is lower than a threshold.

$x_{m,l,c}$: l-th layer inputs

$x'_{m,l,c}$: $(l+1)$-th layer inputs

l: layers

m: each modality sub-network

c: channel

$\mu_{m,l,c}$: mean

$\sigma_{m,l,c}$: standard deviation

$\gamma_{m,l,c}$: trainable scaling factor

$\beta_{m,l,c}$: trainable scaling offset

ϵ: small constant

θ: threshold

Experiments: *Semantic Segmentation*

Dataset: NYUDv2, SUN RGB-D dataset

NYUDv2: 1449 images, 40 classes

SUN RGB-D: 10,335 images, 37 classes

by ResNet pretrained from ImageNet

Table 3: Comparison with SOTA methods on semantic segmentation.

Modality	Approach	Backbone Network	NYUDv2 Pixel Acc. (%)	NYUDv2 Mean Acc. (%)	NYUDv2 Mean IoU (%)	SUN RGB-D Pixel Acc. (%)	SUN RGB-D Mean Acc. (%)	SUN RGB-D Mean IoU (%)
RGB	FCN-32s [33]	VGG16	60.0	42.2	29.2	68.4	41.1	29.0
	RefineNet [31]	ResNet101	73.8	58.8	46.4	80.8	57.3	46.3
	RefineNet [31]	ResNet152	74.4	59.6	47.6	81.1	57.7	47.0
RGB-D	FuseNet [18]	VGG16	68.1	50.4	37.9	76.3	48.3	37.3
	ACNet [21]	ResNet50	-	-	48.3	-	-	48.1
	SSMA [44]	ResNet50	75.2	60.5	48.7	81.0	58.1	45.7
	SSMA [44] †	ResNet101	75.8	62.3	49.6	81.6	60.4	47.9
	CBN [45] †	ResNet101	75.5	61.2	48.9	81.5	59.8	47.4
	3DGNN [36]	ResNet101	-	-	-	-	57.0	45.9
	SCN [30]	ResNet152	-	-	49.6	-	-	50.7
	CFN [29]	ResNet152	-	-	47.7	-	-	48.1
	RDFNet [28]	ResNet101	75.6	62.2	49.1	80.9	59.6	47.2
	RDFNet [28]	ResNet152	76.0	62.8	50.1	81.5	60.1	47.7
	Ours-RefineNet (single-scale)	ResNet101	76.2	62.8	51.1	82.0	60.9	49.6
	Ours-RefineNet	ResNet101	77.2	63.7	51.7	82.8	61.9	50.2
	Ours-RefineNet	ResNet152	77.4	64.8	52.2	83.2	62.5	50.8
	Ours-PSPNet	ResNet152	**77.7**	**65.0**	**52.5**	**83.5**	**63.2**	**51.1**

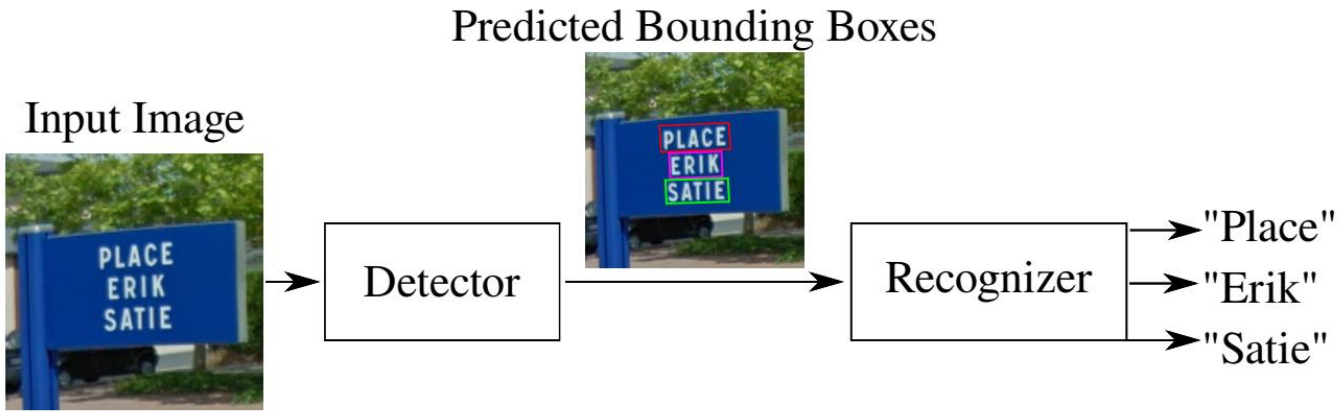
† indicates our implemented results.

Scene Text Detection and Recognition

- 실제 다양한 환경에서 발생하는 이미지에서 텍스트를 탐지하고 이를 컴퓨터 문자로 변환하는 문제

- 이미지 인식 연구의 응용 분야로, Object detection-Semantic segmentation-Sequential image classification 등 알고리즘들이 복합적으로 사용됨

- 이미지 번역, 차량 번호판-이정표-명함 인식, 이미지 검색 등에 실생활에서 다양하게 활용되고 있음



Scene Text Recognition

OCR vs STR

OCR: Optical Character Recognition

종이 문서 등 인쇄된 문자를 인식하는 문제

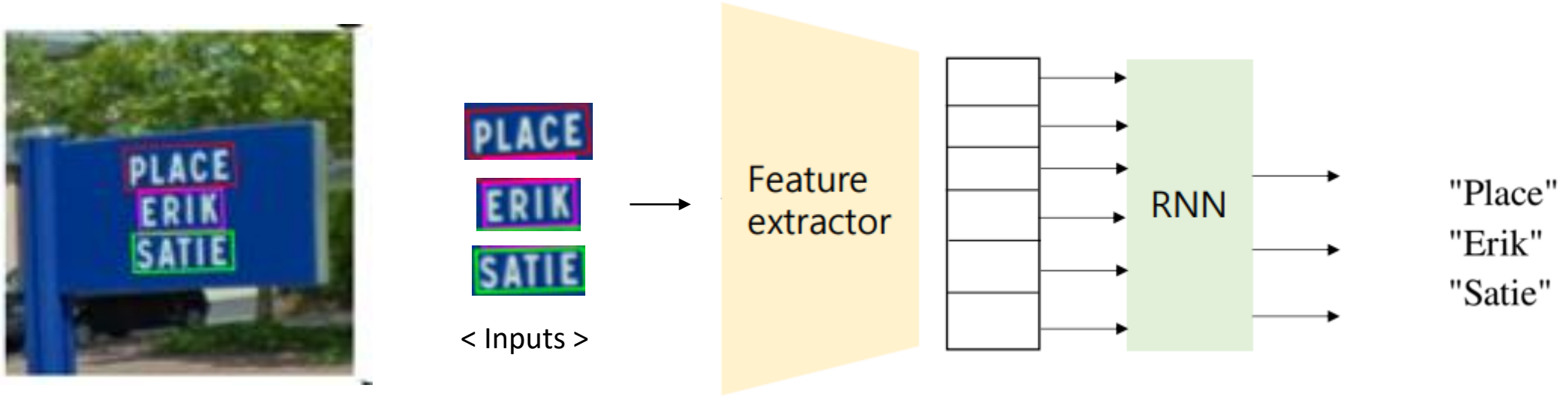
STR: Scene Text Recognition

일상 배경 이미지에서 문자를 인식하는 문제

복잡한 배경과 다양한 글꼴로 구성되어 있어 OCR에 비해 정교한 모델이 필요

Basic of Scene text recognition

- 각 단어 영역이 어떤 문자인지 찾는 Classification 문제
- 단어 영역에 해당하는 이미지로부터 특징 추출 후 Sequential하게 만들어 각 글자의 조합을 찾아가는 방식



Scene Text Recognition Papers

- Convolutional attention networks for scene text recognition

Xie, Hongtao, et al. "Convolutional attention networks for scene text recognition."

ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15.1s (2019): 1-17.

21.10 기준 52회 인용

- FOTS: Fast Oriented Text Spotting with a Unified Network

Liu, Xuebo, et al. "Fots: Fast oriented text spotting with a unified network."

Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

21.10 기준 323회 인용

인공지능 대학원 지원사업(~ 2021.11.30)

- **산학연 3 실적**
  - 비 SCI 게재 1 – *한국지능정보시스템학회 2021 추계학술대회 제출 예정 (~ 21.10.29)*
  - 오픈소스 플랫폼 등록 2 - *11.02 (화) 등록 보고 예정*
    - ✓ 제목: Attention LSTM을 적용한 주가 예측 시스템
    - ✓ 제목: Transformer을 적용한 주가 예측 시스템
- **산학연 5 실적**
  - SW 등록 1 - *"갑상선 안병증 추론을 위한 안와내 주요 구조물 자동 세그멘테이션 프로그램" 등록 완료*

담낭 용종 계획 보고 - 서울대병원

- **신규 데이터 86명(871장)에 대한 데이터 검토 및 전처리 수행** *10.28 (목) 완료 예정*

- **전처리 수행한 신규 데이터와 기존 서울대 데이터를 포함하여 Top-3, Top-5 실험 결과 보고** *11.02(화) 보고 예정*